

# aCGHViewer 2.12 User Guide



© 2005, Roswell Park Cancer Institute

## ***Disclaimer***

The software (aCGHViewer) is copyrighted by Roswell Park Cancer Institute and is released under the LGPL license (<http://www.gnu.org/copyleft/lesser.html>). The software is provided “as is” and Roswell Park Cancer Institute does not assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed.

# Preface

## *How to use this document*

This User Guide can be treated as the operating manual for the aCGHViewer application from Roswell Park Cancer Institute. The rationale and scientific aspects of the program are covered in the Cancer Informatics paper. Any corrections, suggestions for features and bug reports should be sent to ganesh.shankar@roswellpark.org.

Please use the following citation when you utilize this program in your research : Shankar G, Rossi MR, McQuaid DE, Conroy JM, Gaile DP, Cowell JK, Nowak NJ, Liang P. aCGHViewer: a generic visualization tool for aCGH data. **Cancer Informatics** 2006:2 36-43

## Introduction

This informal Guide is written for the Researcher and is meant to provide a “hands-on” tutorial for using aCGHViewer.

The purpose of aCGHViewer is to streamline the analysis of aCGH and other microarray data. The program presents the data in a graphical format for visual inspection and exploration. You can click on any data point of interest to launch a query to the UCSC or NCBI genome browsers. From the genome browser, you have access to any number of inter-related data.

## Downloading and Installation

You will need to install the Java Development Kit (JDK) or Java SDK (version 1.4 or later) from <http://java.sun.com> for aCGHViewer to function. The Java Runtime Environment (JRE) commonly installed on computers is not sufficient.

aCGHViewer is a standalone desktop application and can be downloaded from <http://falcon.roswellpark.org>. The software is Open Source and released under the LGPL ( <http://www.gnu.org/copyleft/lesser.html> ). A copy of the LGPL is located in the CONFIG folder. If you use this program, please cite Shankar G, Rossi MR, McQuaid DE, Conroy JM, Gaile DP, Cowell JK, Nowak NJ, Liang P. aCGHViewer: a generic visualization tool for aCGH data. **Cancer Informatics** 2006:2 36-43.

If you need the source code or want to report bugs, please email ganesh.shankar@roswellpark.org.

The downloaded package is compressed and needs to be uncompressed. After expansion, an application folder is created which contains the aCGHViewer application, a configuration folder named 'CONFIG' and a User Guide. Please ensure that the configuration folder name is in capitals. Place the application folder wherever you want, as long as you have read/write/execute privileges to that location. After the installation of the JDK (see above), the installation is complete and you are ready to use the program. The program can be launched by double clicking the aCGHViewer icon. See the Reference section for launching the application from the command line.

# Data Input

## **Loading data**

aCGHViewer initially checks for the CONFIG folder and displays an error dialog if the folder is not found. After you start the application, a blank window appears with two menu items – File and Tools. Before loading data, ensure that the correct options are selected for your type of analysis. Select Tools>Options and browse through the “Graph”, “Website” and “Organism” tabs and confirm the appropriate choices have been made.

The File menu can be used to load data files. If you select ‘Open’, then a file chooser is displayed allowing you to choose the data file. Multiple files can be chosen at one time. The data input files need to be in tab delimited format with a .txt or .view extension. Files with other extensions can be loaded – you just have to choose “All Files” in the file chooser dialog. The first line of the file should be a header line with titles for each of the columns.

All datasets must contain 4 columns - Probe ID, Chromosome, Position basepair, and value. The probe Id is the identifier that is used to uniquely identify a probe and is used to query the genome browser websites. The chromosome term should be formatted as ‘chr1’, ‘chr2’ etc and represents the chromosome that the probe is located on. The position basepair is a single number that is used to graph the probe along the X axis. This can be the central basepair of a probe. The last essential column lists the value that is the fluorescence intensity of the hybridization. The value may or may not have been normalized or statistically treated. An optional fifth column can be present containing categorized data (see below).

There are three types of datasets that can be loaded – simple, categorized, and combined. By default the simple dataset option is selected when the application is executed for the first time. Before loading data, you should navigate to the Tools>Options menu (see below) and set preferences for your type of analysis.

## **Simple data**

This is the type of dataset that most users will be using. The simple dataset file consists of 4 columns containing information as detailed above. Ensure that Tools>Options>Graph Type is set to “Simple” and aCGHViewer will automatically parse the data.

## **Categorized data**

This type of dataset is the same as the simple dataset except with the addition of a fifth column containing a category flag. You may have categorized data because you statistically processed the data and each probe is now classified as belonging to a particular category. For example, the categories may be 1)unchanged copy number, 2)amplified copy number or 3)decreased copy number. Currently, aCGHViewer supports three flags – “1” for unchanged, “2” for deletions, and “3” for amplifications. Category 1 probes are colored black, category 2 probes are colored green, and category 3 probes are colored red. The association of the category number and color is static at this time, but may be changed in the future. Note that the flagging mechanism can have other semantic meanings depending on the type of experiment being visualized. Ensure that Tools>Options>Graph Type is set to “Simple” and aCGHViewer will automatically parse five column data as categorized data.

## **Combined (overlay) graphs**

This is when you want to graph two sets of data on a single graph. For example, you might want to graph aCGH data and expression data to try and correlate the two. The two datasets are assumed to

be in two different files. You have to select “Combined” as the graph type from the Tools>Options>Graph Type dialog. When loading the files, two successive file dialogs are shown - the first for aCGH (BAC) data, and the second for expression data. After choosing the two files, aCGHViewer will produce a genomic view as before. The expression data is not shown in the genomic view but is shown in the detailed view. Note that the datasets must be of the simple type for combined graphs. Additionally, you are shown an error dialog if you choose to cancel loaded the second file. If you forgot that you had selected the “Combined” graph type, then you need to change it to “Simple”.

*(aCGHViewer 2.05 -- When loading data, you might get an error dialog with words to the effect that the dataset does not contain human or mouse BACs. aCGHViewer is constructed to work with BACs and it checks if the data files contain the string “RP11” or “RP23” as part of the probe identifiers. If it does not find the appropriate string, then it throws an error. This error is legitimate if you have specified a mouse dataset but load a human (“RP11”) dataset or if you have specified a human dataset but load a mouse (“RP23”) dataset. However, if you’re not dealing with BACs at all, then the error is illegitimate and can be ignored. This check is not performed on the Mac.)*

The notes in the previous paragraph do not apply to aCGHViewer 2.12 onwards. Based on user feedback, the error checking has been disabled. However, dialogs are still displayed if the number of columns in the data file is greater than five or lesser than four.

## Options

The options dialog allows you select various settings for the program. The options dialog is located under the Tools menu item. The options dialog is divided into three tabs – Graph, Organism and Website.

### Organism tab – mouse or human

You can select the organism that you’re working on in this tab. Currently, the choice is between mouse and human. You should make this choice before loading the data or a portion of your data may be ignored and an incorrect number of chromosomes will be displayed in the genomic view.

### Website tab – NCBI or UCSC

aCGHViewer launches a query to a genome browser when you click on a data point. The genome browser then returns the results as a web page. You can specify which genome browser ( UCSC or NCBI ) is queried using controls in this tab. You can change the selection after loading the data or in the middle of an analysis. The default setting uses the UCSC genome browser. When “breakpoint zooming”, UCSC is the only genome browser that can be queried.

### Graph tab

The Graph tab is divided into three sections – Graph, Dataset, and Range.

In the Graph section, you can select either “Scatter” or “Line” graph. The default is a scatter graph. The user may choose the format before or after loading the data. If chosen after loading the data, the graph is redrawn when the Options dialog is closed.

In the Dataset section, you can select “Simple” or “Combined” datasets. Choose “Simple” if working with one type of experiment, for e.g. aCGH or expression. Simple graphs can be produced from both simple (4 column) and categorized (5 column) datasets. The dataset type is automatically detected by aCGHViewer. Choose “Combined” if you want to produce an overlay graph with data from

different kinds of experiments, for e.g. aCGH and expression. Overlay graphs can only be produced from 4 column datasets. The default is “Simple”.

In the Y Axis Range section, you can choose the type of Y Axis you want – “Absolute”, “Experiment Relative” or “Chromosome Relative”. These choices determine the scale of the Y Axis. For example, when you choose “Absolute”, you can specify the scale of the Y axis by entering two numbers. The graph is plotted according to that scale while ignoring the range of actual data values. The advantage of this setting is that it allows you to compare results across tumors because they have all been graphed according to the same scale. The disadvantage is that some datapoints may lie outside the scale chosen. The default setting for this section is “Absolute” with a maximum setting of 5 and a minimum setting of -2. Changing the range after the graph is displayed redraws the graph after the Options dialog is closed.

The “Experiment Relative” setting tells the program to calculate the maximum and minimum value within each dataset and use those as the maximum and minimum for the entire set of graphs for one dataset. The advantage of this setting is that no datapoints will be outside the range of the Y axis. The disadvantage is that results between tumors cannot be compared.

The “Chromosome Relative” setting tells the program to calculate the maximum and minimum within each chromosome and use those as the maximum and minimum for each graph. The disadvantage of this choice is that it results in a “noisy” genomic view.

The advantages and disadvantages of each choice is further discussed in the Cancer Informatics paper.

## Genomic view

When a dataset is successfully loaded, the first window to be displayed is the genomic view of the data. Each dataset is loaded in its own tab and multiple datasets can be loaded. Each genomic view consists of individual chromosome panels. Each chromosome panel contains a scatter or line plot of the datapoints in that chromosome. A scatter plot is the default plot type but can be changed in the Tools>Options>Graph dialog. The filename minus the last four characters is used as the tab title. The genomic window can be resized for optimal visualization and allows you to rapidly scan for gains, amplifications or deletions. Clicking on a particular chromosome panel launches a detailed view of just that one chromosome. The genomic view can be printed by choosing File>Print. A landscape orientation can be specified in the print dialog that appears. The best results are achieved if the page margin is set to 0.25 inches on all sides. The file name is printed at the bottom of the printout.

If a blank genomic view appears, then usually, there was a problem with the data loading. See the reference section for tips.

## Detailed view

The detailed view appears when a chromosome panel is clicked in the genomic view. The filename is displayed in the title bar of the detailed view and there is a drop down box at the top of the window. This drop down box allows you to navigate to other chromosomes in the dataset without going back to the genomic view. The detailed view window can be resized and you can zoom into a particular region of the chromosome. Zooming is performed by click-dragging the mouse in a rightward and downward motion over the area of interest. A black rectangle outline is drawn showing the region that will be magnified. When you release the mouse button, the selected region is magnified to fit the

window. You can zoom out by click-dragging in a leftward and upward motion or by using the right-click (control-click on Macs) button. A number of options, including saving the view in graphic (.png) format, are available through the right-click (control-click on Macs) button. In the right-click menu, “Domain Axis” refers to the X axis and the “Range Axis” refers to the Y axis.

You can get additional information about a datapoint by hovering the mouse over it. A tooltip will appear containing the probe ID, cytoband and its y value. The cytoband information is retrieved from the “human\_cytobands.txt” or “mouse\_cytobands.txt” files.

You can also annotate a particular data point by alt-clicking (Windows) on a datapoint. The datapoint is circled, and the Probe Id, cytoband and value information is displayed in red with an arrow pointing to the datapoint. Mac users can perform the same operation using shift-option-click. The caps-lock key can be engaged to make this maneuver easier. On Linux, the alt function can be engaged by clicking the middle button on the mouse.

## **Browser launch**

If you are interested in finding out which genes or other features are located near the probe, then you can click on the datapoint. This will launch a query to either the UCSC or NCBI genome browser using the probe ID as the query term. The default genome browser is UCSC but can be changed in the Tools>Options>Website dialog. If NCBI is the specified website and the datapoint represents a non-BAC probe, and is the second file of a ‘Combined Dataset’ then the query is sent to the Unigene database. A probe is defined to be non-BAC if its identifier does not contain “RP11” or “RP23”.

aCGHViewer launches the system default web browser on Macs and Linux/Unix computers but uses Internet Explorer on Windows. A new browser window is started for each click.

## **Breakpoint zoom**

If you are interested in finding all genes in a particular area of the chromosome, then you can launch a “breakpoint zoom” from the detailed view. You may find a particular region interesting because you suspect it to contain a breakpoint or for some other reason. To launch a “breakpoint zoom”, you need to hold down the shift key while click-dragging the mouse in a rightward and downward direction over the region of interest. The program highlights the selected region in yellow. Please ensure that you are still holding down the shift key when you release the mouse button. The breakpoint zoom is launched only to the UCSC genome browser.

## **Printing**

There are three printing commands available in aCGHViewer – in the right-click menu (control-click for Macs) of a detailed view window, and “Print” and “Print All” from the File menu of the main window. The print command from the detailed view prints the current window to the page. If you have zoomed in to a particular region, then the zoomed view is printed. The “Print” command from the File menu prints the current genomic view. The “Print All” command prints all the currently open genomic panels in a processive fashion. Please be sure to select the “Landscape” orientation in the print dialog as this ensures the best view of the data. For the genomic view, the landscape orientation, with an all-around margin of 0.25 inches usually results in satisfactory results.

If you have problems printing the genomic view in Windows, try to print the view from the main window without resizing it. That is, don’t resize the main window from its default size. Or, experiment

with the page margins and window size. An optimal and robust solution has not been yet discovered by the author.

## Reference

### ***Command line***

The UNIX version of the application is distributed as a .jar file. This file can be launched from the command line with additional flags. By default, aCGHViewer requires at least 128 (aCGHViewer 2.05) or 256 (aCGHViewer 2.12) MB of memory. The increase in memory is necessitated because of larger data files with 100k data points each. If your computer has less, then you can launch aCGHViewer with 'java -jar -Xms64M aCGHViewer.jar '. There are a number of other arguments that can be used and they are described in more detail at [www.java.sun.com](http://www.java.sun.com).

### ***Data format***

The program accepts two kinds of tab-delimited files as input . The files can contain either four or five columns of data. The four column file is used in conjunction with a 'Simple Dataset' setting in Tools>Options>Graph. The five column file is used in conjunction with a 'Combined Dataset' setting in Tools>Options>Graph. aCGHViewer reads the first line and counts the number of elements. Each element is treated as one column. If there are more than five or less than four elements, the program displays an error dialog. If the number of actual data columns do not match the number of heading elements, then the behavior of the program will be erratic. Data lines that contain a Probe Id, but do not contain positional information are discarded. Sometimes, a number of extraneous tabs are inserted by programs. These invisible tabs can throw off the program.

Currently, the five column dataset is used to display categorized data. Each category is indicated by a flag and can be any single alphanumeric character. aCGHViewer only accepts 1,2,and 3 as flags. In aCGH context, all data points categorized as 1 will be colored black, those categorized as 2 will be colored green, and those categorized 3 will be colored red. In a future version, the user will be able to select the number of flags, the associated alphanumeric character and color.

### ***Config files***

A 'CONFIG' folder (directory) should be present in the same directory as the application. This folder contains a number of configuration files that can be manipulated to modify aCGHViewer. Strictly speaking, only the 'log4j.properties' file is essential for the correct functioning of the application. All the other files are optional. If the centromere file is not present, then a vertical line is not drawn on the graphs. If the cytoband file is not present, a 'null' is displayed in the tooltip and annotation.

### ***Cytoband file***

Two files named 'human\_cytobands.txt' and 'mouse\_cytobands.txt' should be present in the 'CONFIG' folder. These files are used to generate the cytoband term in the tooltip. The tab-delimited files are composed of two columns – probe id and cytoband. In this release, aCGHViewer includes the 19k BAC array from the Roswell Park Microarray Facility, the U95E and U133\_Plus\_2 sets from Affymetrix for humans. For Mouse, aCGHViewer supports the 6.5k BAC array from the Roswell Park Microarray Facility, the 430A\_2, Mu11ksub\_b, and MG\_74Cv2 from Affymetrix.

If the probe set you use is not supported, you can append the appropriate data, in the correct format, to the cytoband file and the tooltip should display the correct tooltip. The file name should not be changed. The only limit to an expanding cytoband file is the increased memory requirements. It should be noted that the second column is not limited to cytoband information. Any information, such as a gene symbol or name, can be substituted for the cytoband information. The cytoband file is implemented as a HashMap and does not allow duplicate Probe Ids (keys).

## **Centromere file**

The 'human\_centromeres.txt' file contains the base pair location of the center of each human centromere. These two columns of data are used to draw a vertical line on the graphs to indicate the centromere. If the numbers are changed, then the new vertical line will be drawn accordingly at the modified location.

## **Log properties file**

The config folder also contains a 'log4j.properties' file. This file controls the logging behavior of aCGHViewer. Normally, logging is disabled to prevent performance lags, but can be enabled for debugging purposes. Open the properties file with a text file editor (NOT MS Word) and change the line 'log4j.rootCategory=INFO, FileApp, ConApp' to 'log4j.rootCategory=DEBUG, FileApp, ConApp'. The program will now generate a log file called 'aCGHViewer.log' in the 'CONFIG' folder when executed. The log file can be used for debugging.

## ***Performance***

Performance seems to be determined by three factors – CPU speed, RAM, and the video card. The CPU speed and RAM affect the speed at which the application loads and presents the data, while the presence of a video card relates to the responsiveness of the application. Qualitatively, one user utilizing a computer with Pentium IV, 3.4Gz CPU, 2GB RAM, and 28MB video card to visualize ten 19k datasets experiences an acceptable level of performance. Specifying memory arguments for the JVM also seems to help.

## ***Error Dialogs***

To be done.